Appendix I

Steven M. Ruben
Appl. No. 10/662,429

Recombinant DNA

A Short Course

James D. Watson

COLD SPRING HARBOR LABORATORY

John Tooze

EUROPEAN MOLECULAR BIOLOGY ORGANIZATION

David T. Kurtz

COLD SPRING HARBOR LABORATORY

SCIENTIFIC
AMERICAN
BOOKS

Distributed by

W. H. Freeman and Company

New York

Open Reading Frames in DNA Delineate Protein-Coding Regions

A computer can also be used to analyze a long DNA sequence to determine the location of regions that may code for proteins. The computer is instructed to search for "open reading frames," long stretches of triplet codons that are not interrupted by a translational stop codon. This procedure can be very useful when a cloned DNA fragment is known from, say, some functional assay to contain a certain gene, but when the size of the gene or its location on the fragment is not known. If an open reading frame can be found somewhere in the sequence—especially if the frame has an ATG (the universal translation-initiation codon) near the start—it is very likely that this stretch of sequence is in fact the gene; discovery of an open reading frame does not *prove* the existence of a gene, of course, but it at least delineates an area to home in on. Conversely, the lack of an open reading frame in a stretch of sequence that was thought to contain a gene has been used to determine that some "genes"—chromosomal sequences that hybridize to specific mRNAs—are in fact pseudogenes, nonfunctional relics that arose during the evolution of gene families. Computer searches for open reading frames have even pointed out sequences that code for mRNAs (and probably proteins) that were previously unsuspected. The long terminal repeat (LTR) of mouse mammary tumor virus (Chapter 10) and a stretch of adenovirus DNA, for example, were found to have long open reading frames that have since been found to code for mRNAs. The proteins coded for by these mRNAs have not yet been determined, but no one would have even *looked* for the mRNAs if the open reading frame had not been found.

Leader Sequences at the NH_2-Terminal Ends of Secretory Proteins

DNA sequence analysis reveals that many functional proteins first exist in the form of slightly larger precursors containing some 15 to 25 additional amino acids at their NH_2-terminal ends. Such "leader" (signal) sequences are diagnostic of proteins that move through cellular membranes to function only after they have been secreted from the cells in which they were made (examples of such proteins are insulin, serum albumin, antibodies, and digestive tract enzymes), or after they have been anchored to the outer surface of a cell membrane (the histocompatibility antigens on the cell surface are an example). A majority of the amino acids found in leaders are hydrophobic, and they somehow function to ensure both the attachment of nascent polypeptide chains to appropriate

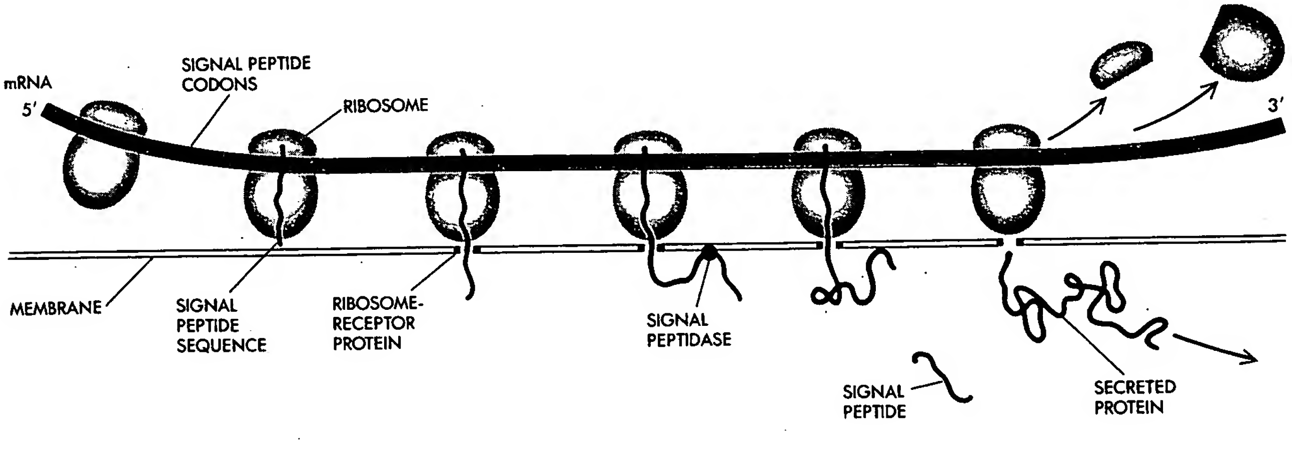

Figure 7-6
Signal sequences. Proteins destined to be secreted from the cell have an N-terminal sequence that is rich in hydrophobic residues. This "signal" sequence binds to the membrane and draws the remainder of the protein through the lipid bilayer. The signal sequence is cleaved off of the protein during this process by an enzyme called signal peptidase.

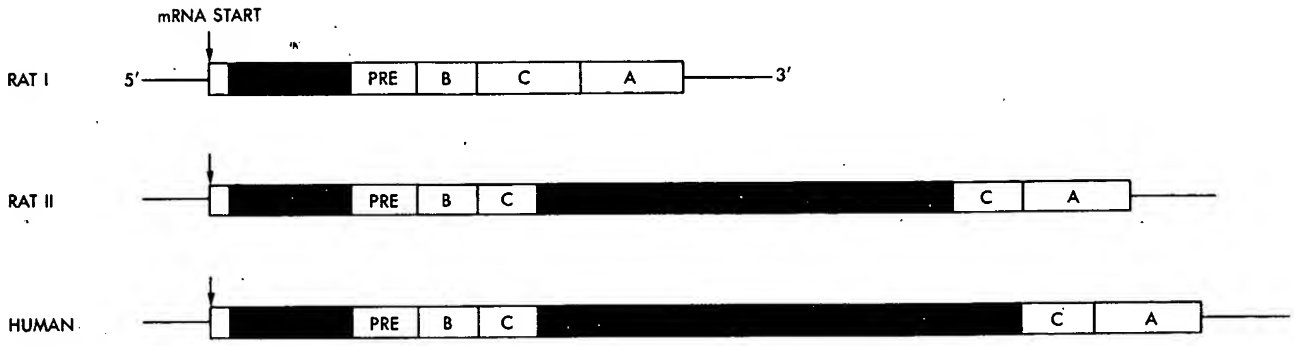

Figure 7-7
A comparison of rat and human insulin genes. Pre, A, B, and C represent the different peptide domains of the proinsulin molecule.

membranes, and the subsequent passage of the chains across the lipid bilayers that characterize all cellular membranes. *In vivo,* leader sequences usually have only a fleeting existence, because they are cleaved off by specific proteolytic enzymes that generate the NH_2-terminal amino acids of the functional secreted products (Figure 7-6).

Introns Sometimes Mark Functional Protein Domains

At first, neither the location nor the number of introns within a given gene made sense. In rats, for example, two closely related genes code for insulin—one gene has only one intron and the other has two. The rat insulin I and rat insulin II genes have introns of almost identical sizes located immediately downstream from the sequences coding for the insulin leader. The second intron of the rat insulin II gene is located within the so-called "C" segment of the insulin protein precursor that is digested away to produce the two-chained structure of mature insulin molecules. Humans have only one insulin gene whose two introns are located in positions similar to those of the rat insulin II gene (Figure 7-7), thus suggesting the descent of rat and human genes from a common ancestor. No obvious functional difference marks the amino acids separated by the second insulin intron, whose location might be accidental.

In hemoglobin, though, the amino acids constituting the special functional domain surrounding the heme group are clearly delineated by an intron from the more distal amino acids. As we describe below, introns within antibody genes are precisely located between functional domains. For this reason, much protein evolution may have been accomplished by genetic recombination events that brought together domains previously located on separate genes. It is conceivable that the long length of many introns helps to ensure that coding sequences are kept intact during genetic crossing over.

Alternative Splicing Pathways Generate Different mRNAs from a Single Gene

RNA splicing can also generate different mRNAs and thus different proteins from one gene, or, more accurately, from one primary transcriptional unit. Differential splicing was first seen in the adenoviruses and then in SV40, in polyoma virus, and in the mRNAs coding for immunoglobulins. A recent example involves the mRNA coding for the hormone calcitonin, a peptide that is normally produced in large amounts in the thyroid gland. Although a large amount of calcitonin mRNA is present in the hypothalamus, very little calcitonin itself is produced there. Instead, another protein that is called "calcitonin-gene-related product" or CGRP, and whose function is still unknown, has been detected. Both calcitonin and CGRP are produced from the same primary transcript by using alternative splicing routes. The routes used produce two different mature mRNAs having a common 5′ end but different 3′ ends: The thyroid